機械翻訳システムの現状と課題

Current situation and problems on machine translation systems

追手門学院大学心理学部教授
井佐原　均

通商産業省工業技術院電子技術総合研究所、郵政省通信総合研究所、独立行政法人情報通信研究機構、豊橋技術科学大学を経て、現職。産業日本語研究会世話人会代表。

1 はじめに

ニューラル機械翻訳によって機械翻訳の実用性は大いに高まった。多くの場面で利用される可能性があり、機械翻訳システムの正確性がますます重要となってきている。本稿ではニューラル機械翻訳の性能向上や課題を実例に沿って述べる。データの重要性や国際標準化の動きに触れ、機械翻訳システムの社会的責任についても考察する。

2 ニューラル機械翻訳の性能向上と残る課題

ニューラル機械翻訳が実用化されたころから筆者が用いてきた例文を図１に示す。上段の日本語を入力した場合の統計翻訳の出力、当時ネット上で使用可能であったニューラル機械翻訳の出力、最近のニューラル機械翻訳の出力を並べた。統計翻訳の訳文がそのままでは使いづらい品質であるのに対し、ニューラル機械翻訳は当時のシステムでも非常に流暢な英文を出力していた。しかし、統計翻訳が「小坪」という地名を「kotsubo」と正しく訳出するのに対し、当時のニューラル翻訳では第１文の訳では「kobosa」と訳出され、第２文では「kosugato」と訳出されている。これはニューラル機械翻訳において低頻度語の処理が難しいこと、また機械翻訳システムが一般に１文単位で翻訳していることが原因である、といった説明をしてきた。

本稿の執筆時点で同じシステムを用いて再度翻訳したところ、図１の最下段に示す訳文となった。初期のニューラル機械翻訳と比べて改善がみられる。「崖崩れ」の訳が「cliff collapse」から「landslide」に変わって

- **逗子市小坪5-1の小坪海岸トンネル鎌倉側で、9月24日0時頃、大きな崖崩れが発生しました。そのため、小坪海岸トンネルや国道134号線における通行に規制がかかっています。**

➢（統計翻訳）Tunnel Kamakura, zushi, kotsubo 5-1 kotsubo coast as 9/24 about a big landslide has occurred. As a result, traffic in kotsubo coast tunnel and route 134, on regulations.

➢（初期のニューラル翻訳）A large cliff collapse occurred around 0 o'clock on September 24th at the Kobosa coast tunnel Kamakura side of Zushi-shi Kobosa 5-1. For that reason, traffic restrictions are imposed on the Kosugato coast tunnel and National Route 134.

➢（最近のニューラル翻訳）A large landslide occurred around midnight on September 24th at the Kamakura side of the Kotsubo coast tunnel at 5-1 Kotsubo, Zushi City. Therefore, traffic on the Kotsubo Kaigan Tunnel and National Highway No. 134 is restricted.

図１　統計翻訳とニューラル翻訳

① そのため、小坪海岸トンネルや国道134号線における通行に規制がかかっています。

➢ Therefore, traffic on the Kotsubo Kaigan Tunnel and National Highway No. 134 is restricted.

② そのため、小坪海岸トンネルや国道134号線における走行に規制がかかっています。

➢ Therefore, there are restrictions on driving in the Kotsubo coast tunnel and National Highway No. 134.

③ そのため、小坪海岸トンネルにおける通行に規制がかかっています。

➢ Therefore, traffic in the Kotsubo coast tunnel is restricted.

④ そのため、小坪海岸トンネルにおける走行に規制がかかっています。

➢ Therefore, driving in the Kotsubo coastal tunnel is restricted.

図２　訳の揺れ

いる。地名に関わる訳出も向上が見られる。「小坪」も「Kotsubo」と正しく訳出されている。ニューラル機械翻訳がデータ、アルゴリズム、計算パワーにより着実に性能向上が進められていることが分かる。

しかしながら、日本語原文の第１文の「小坪海岸トンネル」が「Kotsubo coast tunnel」と訳されているのに対し、第２文では「Kotsubo Kaigan Tunnel」と訳されるなど、課題が残る。この辺りの挙動を確かめるために、図１の日本語原文の第２文を元に変更を加えた文をいくつか最新のニューラル機械翻訳で翻訳してみた。結果を図２に示す。①は図１の日本語原文のままである。この文に対し、「通行」を「走行」に変えただけで（②）、訳文の構造が変わり、さらには「小坪海岸トンネル」の訳語が「Kotsubo Kaigan Tunnel」から「Kotsubo coast tunnel」に変わっている。①から「国道 134 号線」に関する部分を外すと（③）、訳文の構造は変わらないが「on」が「in」になり、先ほどと同様に「小坪海岸トンネル」の訳語が変わる。③に対し、さらに「通行」を「走行」に変えると（④）、文の構造は変わらないが、「小坪海岸トンネル」が「Kotsubo coastal tunnel」と変わる。④は②から「国道 134 号線」に関する部分を外したとも考えられる。

3 ブラックボックス化するシステムとデータ

前章のようなニューラル機械翻訳の動作は、その理由が分かりづらく、改善が難しい。旧来の機械翻訳システムのように文法や辞書といった明示的な規則を用いているのではないため、入力と出力の間で何が起こっているかがブラックボックス化している。機械翻訳システムをはじめとする自然言語処理システムがルールベースであったときには、システムの個々の誤りをルールの追加や修正で正していくことが可能であった。しかしながら、ニューラル機械翻訳は大量の正解データを用いた学習によって精度を向上する。入力と出力を繋ぐ部分（ニューラルネット）はブラックボックスであり、個々の誤りを修正するのは難しい。

ニューラル機械翻訳をはじめとする人工知能システムが人間にとって完璧なものでなく、人間の補助として用いるべきものであるとするならば、人間が最終判断をするために必要な情報を得ることが必要である。ブラックボックスではない、理由を説明できるシステムが求められよう。

ニューラル機械翻訳は、大規模なデータからの学習によって実現されているため、学習データの質が悪ければ、良い出力を得ることはできない。機械翻訳システムに限らず人工知能システムの性能を担保するためには、学習のアルゴリズムの適切性はもちろん、データやアノテーション（情報付与）の適切性を保証することも重要である。従来のシステム開発は、人間が生のデータを見て分析し、必要な処理を定義する。それを実装（プログラミング）して、製品に組み込むという手順で行われた。それぞれのステップでの責任が明確であり、トラブル時の責任範囲が確定できた。一方、深層学習などの機械学習による人工知能システムの開発では、データを収集し、アノテーションし、そのデータを用いて学習を行い、得られたモデルを製品に組み込む。この場合、学習部分がブラックボックス化していることに加えて、そもそものデータの品質（由来など）が保証できるか、データへのアノテーションが適切か、といった点を考慮することが

必要である。利用したデータが適切か（たとえば、話し言葉の処理システムの学習に書き言葉のデータを使っていないか）、アノテーションにおいて、その体系は適切か、アノテーションの精度は十分に高いか、といった点が重要となるが、これらを保証したデータは限定されよう。

4 データと国際標準化の動き

前章で述べたように機械翻訳をはじめとする人工知能システムの実用化には言語資源の作成やアノテーションにおける質の保証が重要である。国際標準化機構（ISO: International Organization for Standardization）においてもコーパスへのアノテーションプロジェクトにおける質保証を対象に標準化が始まっている。ISO には多くの専門委員会（TC: Technical Committee）が存在するが、その中で TC37 Language and terminology は専門用語、言語、内容の情報資源を対象としている。TC37 の分科委員会（SC: Sub Committee）の一つである SC4 Language resource management において、Corpus Annotation Project Management の新規提案が行われている。コーパスアノテーションプロジェクトの目標は、限られたリソース環境内でアノテーション仕様に従って誤りのない成果物を実現することである。標準化の提案は誤りのないアノテーション付きコーパスを効果的かつ効率的に構築するための推奨事項を提供することを目的とした一連の標準として提案されている。コーパスアノテーションプロジェクトの開始段階から終了段階までを対象とするが、開始前の計画段階は対象としていない。現時点では、図 3 に示す三つの標準化が提案されている。

筆者は ISO TC37 の国内委員会の委員長であり、上記の提案のうち、PWI 24635-2 と PWI 24635-3 の co-project leader を務めている。

5 機械翻訳と責任

機械翻訳の性能が向上し、その出力がそのまま社会で利用されるようになると、誤訳に対する責任が問題となる。特に医療や法務の場面において機械翻訳を用いた通訳システムが使われるような場合には大きな問題となろう。

人工知能と社会との関わりを語るときに、自動車の自動運転が例に挙げられることがある。技術的には日々進歩し、市販されている乗用車にも自動ブレーキや車線の自動変更機能が付いたものが出ている。さらに進んで人手の介入のない完全自動の自動車が実現した場合の問題として述べられるのが、いわゆるトロッコ問題である。自動運転車がそのまま進めば 5 人を犠牲にする状況で進路を変えて 1 人を犠牲にすることは是かというものである。しかしこのような選択は人間が運転している場合でも起こりうる。実際の社会では事故が起こった後で法律上の判断が行われる。人間の判断を事後に評価しており、運転手がその判断に至った元になる情報や規則は事前に明示的には表現されていない。しかし人工知能システムが運転する場合には、ある状況においてどのような判断をするかは、データと学習機構によって事前に決まっており、シミュレーションによって検証できる。ある場面において、ある決断をすることが適切か、それを誰が判断するのかが議論となる。この議論においても人

- ISO PWI 24635-1: Language resource management – Corpus Annotation Project Management – Part 1: Core Model （コーパスアノテーションで検討すべき事項，コーパスアノテーションプロジェクトの手順，プロジェクトの編成，作業パッケージ，およびコーパスアノテーションプロ ジェクトの規模，複雑さ，期間に関係なく適用できるタスクを含む基本原則に関する標準．）
- ISO PWI 24635-2: Language resource management – Corpus Annotation Project Management – Part 2: Training Model （プロジェクト参加者をトレーニングし，プロジェクトを実行する能力を維持するための基本原則に関する標準．）
- ISO PWI 24635-3: Language resource management – Corpus Annotation Project Management – Part 3: Validation Model （アノテーションの仕様に従って誤りのないアノテーションを実現する成果物の品質管理の基本原則に関する標準．）

図 3 ISO におけるコーパスアノテーションプロジェクトの標準化

工知能システムの透明性（説明能力）が重要である。

機械翻訳の場合も、ある入力に対してどのように出力されるかは事前に決まっており、誤訳が起こる場合も決まっている。そのようなシステムを人命にかかわるような場面で用いるかどうかは難しい。法廷に機械翻訳が入ることによって、言語障壁によって不当な判決を下されていた多くの人が助かるとしても、一人でも冤罪となる人が出るとしたら、我々はどのようにするべきであろうか。何が正しいかは哲学の問題となり、人工知能ではなく人間が判断すべき事柄であろう。

機械翻訳のように人間の知的活動を支援あるいは置換するシステムには自動運転とは異なる点がある。自動運転の場合は、車道と歩道の分離や立体交差化を行い、すべての車が制限速度を守って自動走行するようになれば、トロッコ問題のような事態を発生させないようにすることが可能であろう。しかし翻訳などの高度な判断を要するタスクにおいては、このような安全な環境を作ることは想定しがたい。ここでもまた哲学の領域となるのかもしれない。

6 おわりに

コンピュータはゲームや物理現象などのようにルールが定まっている問題には、そのルールが明示的であるかないかに関わらず、人間以上に対応できることが多い。しかしながら、言葉や心の処理では、まだまだ人間と同じレベルに届かない。ひとりの人間が様々な場面で適切にコミュニケーションできるように、様々な場面で人間よりうまくコミュニケーションができる人工知能の実現はまだまだ先であろう。

機械翻訳の最終的な目標は、伝えたい情報（意図）を伝える翻訳の実現であろう。広告を翻訳するときには文を訳すことだけを目指したのでは不十分であり、目標言語（翻訳先言語）の読み手がその商品を買いたくなることが必要である。機械翻訳の守備範囲を超えているかもしれないが、翻訳の本来の目標はそんなところにあるのかもしれない。